Journal of Computational Intelligence in Bioinformatics
ISSN 0973-385X Volume 5, Number 2 (2012) pp. 141-148

http://www.ripublication.com/jcib.htm

Machine Learning Simulation Model for Prediction and Classification of Subcellular Localization of HIV Apoptosis Proteins by Amino acid Composition

Anubha Dubey[1], Kumud Pant[1] and Dr. Usha Chouhan[2]

[1]Ph.D. Scholar, [2]Assistant Professor,
Department of Bioinformatics, MANIT, Bhopal-462051, M.P., India
E-mail: anubhadubey@rediffmail.com

Abstract

Protein (or in general, proteome) Analysis Subcellular Localization Prediction is a process (usually through the use of web-based software) of predicting the location or destination of a protein within the cell using only the protein sequence as its inputs. Proteins are then likened to letters with proper address and stamps to deliver it on the proper destination. Since the proteins should have proper address to ensure its delivery to the proper localization. The destination of various protein sequences is predicted by the subcellular localization prediction servers. Hence a machine learning simulation model is developed to predict and classify HIV apoptosis proteins subcellular localization sites by their amino acid composition. Of the various predictions software's used Eukaryotic Mploc predicts better results mitochondria with accuracy of 99.1304%, Subloc shows better results with mitochondria with accuracy of 90%, and Virus Ploc shows better results with extracellular space with accuracy of 98.889%.

Introduction

Apoptosis, also known as programmed cell death, is characterized by specific morphologic and biochemical properties [1]. Apoptosis proteins play a central role in development and homeostasis of an organism [2]. The function of an apoptosis protein is closely correlated with its subcellular location, so it is very important to gain the information about the subcellular location of apoptosis proteins [3]. Although subcellular location of unknown proteins can be determined by experimental methods, they are both time-consuming and expensive. Therefore, it is very urgent to develop an accurate and reliable prediction method for apoptosis protein subcellular location.

The progression of the human immunodeficiency virus infection to AIDS is primarily due to the depletion of CD4+ T-helper lymphocytes, which leads to a compromised immune system. One of the mechanisms by which T-helper cells are depleted is apoptosis, which results from a series of biochemical pathways [15].

1. HIV enzyme cause cell death, but primes the cell for apoptosis should the appropriate signal be received. In parallel, these enzymes activate pro-apoptotic *procaspase-8*, which does directly activate the mitochondrial events of apoptosis.
2. HIV may increase the level of cellular proteins which prompt Fas-mediated apoptosis.
3. HIV proteins decrease the amount of CD4 glycoprotein marker present on the cell membrane.
4. Released viral particles and proteins present in extracellular fluid are able to induce apoptosis in nearby "bystander" T helper cells.
5. HIV decreases the production of molecules involved in marking the cell for apoptosis, giving the virus time to replicate and continue releasing apoptotic agents and virions into the surrounding tissue.
6. The infected CD4+ cell may also receive the death signal from a cytotoxic T cell.

Cells may also die as a direct consequence of viral infection. HIV-1 expression induces tubular cell G2/M arrest and apoptosis [20]. Actually; many efforts have been made for protein subcellular location: Emanuelsson et al. [4] used N-terminal sequence as an input into two layers of artificial neural networks. Zhou and doctor attempted to identify four kinds of subcellular locations of 98 apoptosis proteins based on amino acid composition by means of the covariant discriminates function [5]. Because of the information absence of sequence order in fact, some new protein features were proposed in order to incorporate sequence order effects of proteins. Chou proposed the concept “pseudo amino acid compositions”[6]. Chou and Cai [7, 8] developed an accurate method integrating the pseudo amino acid compositions, the function domain composition and the information of gene ontology. Feng proposed a new representation of unified attribute vector; all of proteins have their representative points on the surface of the 20-D globe [9]. Shao et al used complexity measure factor to predict protein subcellular location [10], Zhang et al predicted protein homo-oligomer types by pseudo amino acid composition which used an improved feature extraction and Naive Bayes feature fusion [11]. Bulashevska and Eils predicted the four kinds of subcellular locations of the same datasets by using hierarchical ensemble of Bayesian classifiers based on Markov chains [12] and so on. Recently, Chen and Li utilized the measure of diversity and increment of diversity to predict the subcellular location of apoptosis proteins [13, 14].In the central dogma of molecular biology, the major copy of the message for protein synthesis is contained in the DNA. The message is transcribed into mRNA and then transported to the cytosol where translation of the copies of mRNA into proteins happens. Several post-translational modification processes happen before the cell could actually use the modified protein. This protein is then transported to a particular organelle (intracellular) or outside the

cell (extracellular) where the protein is needed. Proteins are then likened to letters with proper address and stamps to deliver it on the proper destination.

In this paper, the amino acid composition of a protein sequences are used to construct the model, and use Eukaryotic MPloc [17], Virus ploc [18],Subloc [19] softwares to predict the subcellular location of HIV apoptosis proteins. Amino acids are critical to life, and have a variety of roles in metabolism. One particularly important function is as the building blocks of proteins, which are linear chains of amino acids. Every protein is chemically defined by this primary structure, its unique sequence of amino acid residues, which in turn define the three-dimensional structure of the protein. Just as the letters of the alphabet can be combined to form an almost endless variety of words, amino acids can be linked together in varying sequences to form a vast variety of proteins [16]. Due to their importance a machine learning simulation model is being developed to classify and predict subcellular localization of HIV apoptosis proteins [21, 22, and 23].

Methodology

Data set

To achieve our goal and develop our methodology we obtained the dataset from Swissprot/Uniprot databank of Expasy server [24]. The following two data sets were used.

Dataset 1: It consisted of all the subcellular locations of HIV proteins predicted from Eukaryotic MPloc. The subcellular positions obtained are cell membrane, extracellular, cytoplasm, secreted proteins, nucleus and mitochondria .All the entries marked as fragments were not included in the dataset. The total instances were 290.The 290 were positive belonging to Eukaryotic MPloc and 290 were negative instances belonging to other enzymatic group. Table 1 shows individual position of subcellular location of HIV proteins. The predicted sites with their number of proteins are as- cell membrane=115, extracellular=6, cytoskeleton=8, cytoplasm=27, secreted proteins=22, Nucleus=91, Mitochondria=21.

Dataset 2: It consisted of all the subcellular locations of HIV proteins predicted from Virus Ploc. The subcellular positions obtained are plasma membrane, extracellular, cytoplasm, nucleus and mitochondria .All the entries marked as fragments were not included in the dataset. The total instances were 268. The 268 were positive belonging to Virus Ploc and 268 were negative instances belonging to other enzymatic group. The predicted sites with number of proteins are as: Plasma membrane-165, Nucleus-94, Mitochondria-1, Cytoplasm-7, and Extracellular-1.

Dataset 3: It consisted of all the subcellular locations of HIV proteins predicted from Sub loc software. The subcellular positions obtained are 253 extracellular, cytoplasm, nucleus and mitochondria. All the entries marked as fragments were not included in the dataset. The total instances were 253. The 253 were positive belonging to Sub loc and 253 were negative instances belonging to other enzymatic group. The predicted

sites with number of proteins are as: Extracellular-94, Cytoplasm-112, Nucleus-29, and Mitochondria-18.

Dataset 4: It consisted of all the subcellular locations of HIV proteins predicted from Eukaryotic MPloc, Virus ploc and Sub loc software. The total instances were 806. The 806 were positive belonging to all the mentioned softwares and 806 were negative instances belonging to other enzymatic group.

Support vector machine (Binary classification)

SVM is a supervised machine learning method which is based on the statistical learning theory [25,26,27,28]. When used as a binary classifier, an SVM will construct a hyperplane, which acts as the decision surface between the two classes. This is achieved by maximizing the margin of separation between the hyperplane and those points nearest to it. The SVMs were implemented using freely downloadable software, libSVM [26]. In this software there is a facility to define parameters and choose among various inbuilt kernals. They can be radial basis function (RBF) or a polynomial kernel (of given degree), linear, sigmoid.

SVM software; LIBSVM

Simulations were performed using LIBSVM version 2.89 (a freely available software package) [26]. For our study RBF Kernel was found to be the best. The SVM training was carried out by the optimization of the value of the regularization parameter and the value of RBF kernel parameter.

Amino Acid Composition

Previously, this parameter has been used for predicting the subcellular localization of proteins [10]. The amino acid composition is the fraction of each amino acid type within a protein.

The fractions of all 20 natural amino acids were calculated by using Equation 1,

$$= \frac{Total\ Number\ of\ amino\ acid\ i}{Total\ number\ of\ amino\ acids\ in\ a\ protein}$$

Polycomp

The input vector of 450 was generated directly in the format of SVM by software Polycomp developed under Department of Bioinformatics, MANIT, Bhopal, India [29]. This software generates data which can be directly fed into the classifier hence saving valuable time and energy needed for formatting the hybrid.

Evaluation of Performance

The performance of our classifier was judged by 5 fold cross validation. The LibSVM provides a parameter selection tool using the RBF kernel: cross validation via grid search. A grid search was performed on C and Gamma using an inbuilt module of libSVM tools on Dataset 1, Dataset 2, Dataset 3 and Dataset 4 as shown in Figure1, Figure2, Figure3, and Figure4. Here pairs of C and Gamma are tried and the one with the best cross validation accuracy.

Result & Discussion

Protein Localization Prediction is important to predict the possible location or destination of a particular protein. It should be noted that this is still a prediction. An actual wet-lab procedure is preferred to do actual analysis of the protein's location. Since the prediction servers present possible location of the protein, then it also gives us insights of the possible function of the protein sequence. Since Eukaryotic MPloc, Subloc and Virus ploc are some of the software which predicts better results as shown in table 1, table2 and table2. Table 1 shows the prediction of subcellular localization of HIV proteins with its accuracy, c and gamma values.

Table 1: shows Eukaryotic MPloc results:

SNo.	Protein subcel localization sites	Number of proteins	Accuracy	c	g
1.	Cytoplasm	7	98.889%	-	-
2.	Nucleus	91	96.667%	-	-
3.	Plasma Membrane	115	99.900%	-	-
4.	Mitochondria	21	99.1304%	1.0	0.45
5.	Extracellular	6	96.2104%	-	-
6.	Cytoskeleton	8	-	-	-
7.	Secreted proteins	22	98.113%	-	-
8.	All protein subcel prediction sites	290	98.1865%	-	-

Table 2: shows Subloc results:

SNo.	Protein subcel localization sites	Number of proteins	Accuracy	c	g
1.	Plasma membrane	165	86%	-	-
2.	Nucleus	94	81.3187%	0.03125	0.0078125
3.	Cytoplasm	7	82.9167%	-	-
4.	Mitochondria	1	90%	2.0	0.5
5.	Extracellular	1	87.9167%	0.125	0.125
6.	All protein subcel prediction sites	268	98.604%	0.5	0.5

Table 3: Shows Virus ploc results:

SNo.	Protein subcel localization sites	Number of proteins	Accuracy	c	g
1.	Extra cellular	94	98.889%	0.03125	0.0071825
2.	Nucleus	29	94.9153%	0.5	0.125
3.	Cytoplasm	112	94.382%	-	-
4.	Mitochondria	18	99.2308%	-	-
5.	All protein subcel prediction sites	253	96.2712%	0.125	0.5

Table 4: shows all positives of Eukaryotic MPloc, Subloc and Virusploc.

SNo.	Protein subcel localization softwares	Accuracy	c	g
1.	Eukaryotic MPloc,Virusploc and Sub loc	98.0964%	-	-

The accuracy of an SVM model is largely dependent on the selection of the model parameters. Two methods for finding optimal parameter values, a grid search and a pattern search. A grid search tries values of each parameter across the specified search range using geometric steps. A pattern search (also known as a “compass search” or a “line search”) starts at the center of the search range and makes trial steps in each direction for each parameter. If the fit of the model improves, the search center moves to the new point and the process is repeated. Here in this prediction model grid search is used. Parameters c and gamma (g) is used to predict better accuracies of the softwares used. As shown in all the softwares used for prediction Eukaryotic MPloc predicts better prediction sites for protein subcellular localization. The accuracies of predicted sites are shown in table1.Eukaryotic MPloc shows better results with mitochondria with accuracy of 99.1304%,c=1.0, and g=0.45. Subloc shows better results with mitochondria with accuracy of 90%, c=2.0, g=0.5 (table 2) and Virus Ploc shows better results with extracellular space with accuracy of 98.889%, c=0.03125, and g= 0.0071825 (table 3). Table 4 shows that dataset 4 of all positive values of all the softwares used predicted better prediction sites for protein subcellular localization. The blank spaces of c and g values shows that the SVM could not predict better results with that dataset as the data is not sufficient to predict c and g values or it might be possible that the data is overlapping.

Conclusion

In this paper, we have proposed a method for predicting protein subcellular localization (PSL) for HIV based on amino acid composition. Experiment results show that the SVM model achieves high prediction accuracy for all the data sets, thus supporting the assumption that biological features derived from HIV could significantly improve the performance. With more prediction sites for software used the accuracy of SVM model is further improved, could also be a useful indicator for inferring PSL. The results suggest that the performance could be overestimated if redundant sequences are considered. In the assessment of the evaluation data sets. The proposed method can be used in large-scale analyses of proteomes and is freely available for public use as the database is developed. There are still some challenges to be addressed in PSL prediction. In our work, we only consider proteins with single localization sites. However, proteins with multiple localization sites are not a rarity, especially in higher order species. In our future development, we will consider those proteins localized to multiple compartments. In addition, better accuracy and coverage are needed, particularly for several poorly predicted localization sites. We will also extend our method to combine more biological features, analyze multiple

compartment proteins, and incorporate proteins of more species, including those of humans.

Acknowledgement

The authors are highly thankful to the Department of Biotechnology, New Delhi, India and M.P. Council of Science and Technology M.P., Bhopal, India for providing support in the form of Bioinformatics infrastructure facility to carry out the research work.

References

[1] Wyllie, A.H., Kerr, J.F., Currie, A.R. "Cell death: the significance of apoptosis". Int. Rev. Cytol. Vol. 68, 1980, pp.251–306.

[2] Raff, M. Cell suicide for beginners. Nature. Vol. 396, No. 3707, 1998, pp.119–122.

[3] Suzuki, M., Youle, R.J., Tjandra, N. "Structure of Bax: coregulation of dimmer formation and intracellular location". Cell. Vol. 103, No. 4, 2000, pp. 645–654.

[4] Emanuelsson O, Nielsen H, Brunak S, Heijne G. "Predicting subcellular localization of proteins based on their N-terminal amino acid sequence". J Mol Biol Vol. 300, No. 4, 2000, pp. 1005–1016

[5] Zhou, G.P., Doctor, K. "Subcellular location prediction of apoptosis proteins". Proteins: Struct. Func. Genet. Vol. 50, No. 2, 2003, pp. 44–48.

[6] Chou KC. "Prediction of protein cellular attributes using pseudoamino acid composition". Proteins: Struct. Funct. Genet. Vol. 43, No. 3, 2001, pp.246–255.

[7] Chou KC, Cai YD. "Using functional domain composition and support vector machines for prediction of protein subcellular location". J Biol Chem, Vol. 227, No. 48, 2002, pp.45765–45769

[8] Chou KC, Cai YD. "Predicting subcellular localization of proteins by hybridizing functional domain composition Vol. 91, No. 3, 2004, pp.1197–1203

[9] Feng, Z.P. "Prediction of the subcellular location of prokaryotic proteins based on a new representation of the amino acid composition". Biopolymers, Vol. 58, No. 4, 2001, pp. 491–499.

[10] Xiao, X., Shao, S., Ding, Y., Huang, Z., Huang, Y., Chou, K.C. "Using complexity measure factor to predict protein subcellular location". Amino Acids, Vol. 28, No. 1 2005, pp. 5761.

[11] Zhang, S.W., Pan, Q., Zhang, H.C., Shao, Z.C., Shi, and J.Y. "Prediction of protein homo-oligomer types by pseudo amino acid composition: approached with an improved feature extraction and Naive Bayes feature fusion". Amino Acids, Vol. 30, No. 4, 2006, pp.461468.

[12] Bulashevska A, Eils R. "Predicting protein subcellular locations using hierarchical ensemble of Bayesian classifiers based on Markov chains". BMC Bioinformatics, Vol. 7, No. 1, 2006, pp. 298

[13] Chen Y L, Li Q Z. “Prediction of the subcellular location of apoptosis proteins’. Journal of Theoretical Biology, Vol. 245, No. 4, 2007, pp. 775–783
[14] Chen Y L, Li Q Z, Yang K L, Fan G L. “Predicting of the subcellular location of apoptosis proteins using the algorithm of the increment of diversity combined with support vector machine”. Acta Biophysica Sinica, Vol. 23, No. 3, 2007, pp.192-197.
[15] Judie B. Alim onti, T. Blake Ball, Keith R. Fowke (2003). "Mechanisms of CD4+ T lymphocyte cell death in human immunodeficiency virus infection and AIDS". J Gen Virology 84 (84): 1649–61. doi:10.1099/vir.0.19110-0. PMID 12810858
[16] The Structures of Life". National Institute of General Medical Sciences. http://publications.nigms.nih.gov/structlife/chapter1.html. Retrieved 2008-05-20.
[17] Chou KC, Shen HB (2007) Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites. Journal of Proteome Research 6: 1728–1734. http://www.csbio.sjtu.edu.cn/bioinf/virus
[18] http://www.csbio.sjtu.edu.cn/bioinf/virus
[19] http://www.bioinfo.tsinghua.edu.cn/dbsubloc.html.
[20] Vashistha H, Husain M, Kumar D, Yadav A, Arora S, Singhal PC. (2008)Ren Fail. 2008; 30(6):655-64.
[21] A.Dubey, B.Pant and Neeru Adlakha,”SVM Model for Amino Acid Composition based Classification of HIV1 Groups”. IEEE digital library published.
[22] A.Dubey, B.Pant and Usha Chouhan,” SVM Model for Classification of Structural and Regulatory Proteins of HIV1 and HIV2 Journal of Advanced Bioinformatics Applications and Research ISSN 0976-2604 Vol 2, Issue 1, 2011, pp 84-88
[23] A.Dubey, B.Pant and Usha Chouhan,” Machine learning model for HIV1 and HIV2 enzyme secondary structure classification “J. Comput. Method. Mol. Design, 2011, 1 (2): 1-8
[24] www.uniprot.org
[25] C.C-Cheng and C-J Lin, “LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/clin/libSVM.2001
[26] V.Vapnik,”The nature of statistical learning theory,”Springer” 1995.
[27] M.Wang, J.Yang,G.P.Liu,Z.J.Xu,K.C.Chou,”Weighted support vector machines for predicting membrane protein types based on pseudo amino acid composition, “Protein Eng.Des.Sel,vol.17,pp. 509-516,2004.
[28] Lavanya et al, Bioinformation 5(5): 227 (2010)